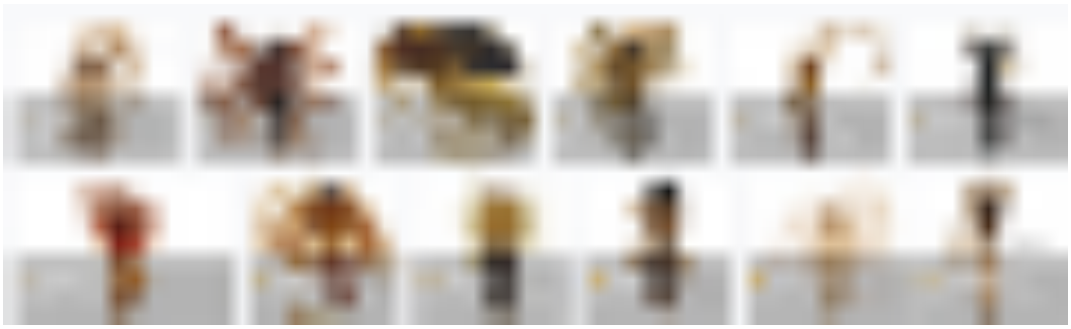
medium.com

Towards a digital natural history museum

Roderic Page

11-14 minutes



These notes are the result of a few events I've been involved in the last couple of months, including TDWG 2017 in Ottawa, a thesis defence in Paris, and a meeting of the Science Advisory Board of the Natural History Museum in London. For my own benefit if no one else's, I want to

sketch out some (less than coherent) ideas for how a natural history museum becomes truly digital.

Background

The digital world poses several challenges for a museum. In terms of volume of biodiversity data, museums are already well behind two major trends, observations from citizen science and genomics. The majority of records in GBIF are observations, and genomics databases are growing exponentially, through older initiatives such as barcoding, and newer methods such as environmental genomics. While natural history collections contain an estimated 109specimens or “lots” [1], less than a few percent of that has been digitised, and it is not obvious that massive progress in increasing this percentage will be made any time soon.

Furthermore, for citizen science and genomics it is not only the amount of data but the network effects that are possible with that data that make it so powerful. Network effects arise when the value of something increases as more people use it (the classic example is the telephone network). In the case of citizen science, apart from the obvious social network that can form around a particular taxon (e.g., birds), there are network effects from having a large number of identified observations. iNaturalist is using machine learning to suggest identifications of photos taken by members. The more members join and add photos and identifications, the more reliable the machine identifications become, which in turn makes it more desirable to join the network. Genomics

data also shows network effects. In effect, a DNA sequence is useless without other sequences to compare it with (it is no accident that the paper describing BLAST is one of the most highly cited in biology). The more sequences a genomics database has the more useful it is.

For museums the explosion of citizen science and genomics begs the question “is there any museum data that can show similar network effects”? We should also ask whether there will be an order of magnitude increase in digitisation of specimens in the near future. If not, then one could argue that museums are going to struggle to remain digitally relevant if they remain minority biodiversity data providers. Being part of organisations such as GBIF certainly helps, but GBIF doesn’t (yet) offer much in the way of network effects.

Users

We could divide the users of museums into three distinct (but overlapping) communities. These are:

1. Scientists
2. Visitors
3. Staff

Scientists make use of research and data generated by the museum. If the museum doesn’t support science (both inside and outside the museum) then the rationale for the collections (and associated staff) evaporates. Hence, digitisation must support scientific research.

Visitors in this sense means both physical and online visitors. Online visitors will have a purely digital experience, but in person visitors can have both physical and digital experiences.

In many ways the most neglected category is the museum staff. Perhaps best way to make progress towards a digital museum is having the staff committed to that vision, and this means digitisation should wherever possible make their work easier. In many organisations going digital means a difficult transition period of digitising material, dealing with crappy software that makes their lives worse, and a lack of obvious tangible benefits (digitisation for digitisation's sake). Hence outcomes that deliver benefits to people doing actual work should be prioritised. This is another way of saying that museums need to operate as "platforms", the best way to ensure that external scientists will use the museums digital services is if the research of the museum's own staff depends on those services.

Some things to do

For each idea I sketch a "vision", some ways to get there, what I think the current reality is (and, let's be honest, what I expect it to still be like in 10 years time).

Vision: Anyone with an image of an organism can get a answer to the question "what is this?"

Task: Image the collection in 2D and 3D. Computers can now "see", and can accomplish tasks such as identify species and traits (such as the presence of disease [2]) from

images. This ability is based on machine learning from large numbers of images. The museum could contribute to this by imaging as many specimens as possible. For example, a library of butterfly photos could significantly increase the accuracy of identifications by tools such as iNaturalist. Creating 3D models of specimens could generate vast numbers of training images [3] to further improve the accuracy of identifications. The museum could aim to provide identifications for the majority of species likely to be encountered/photographed by its users and other citizen scientists.

Reality: Imaging is unlikely to be driven by identification and machine learning, beiggest use is to provide eye-catching images for museum publicity.

Who can help: iNaturalist has experience with machine learning. More and more of research is appearing on image recognition, deep learning, and species identification.

Vision: Anyone with a DNA sequence can get a answer to the question “what is this?”

Task: DNA sequence the collection, focussing first on specimens that (a) have been identified and (b) represent taxonomic groups that are dominated by “dark taxa” in GenBank. Many sequences being added to GenBank are unidentified and hence unnamed. These will only become named (and hence potentially connected to more information) if we have sequences from identified material of those species (or close relatives). Often discussions of sequences focus on doing the type specimens. While this

satisfies the desire to pin a name to a sequence in the most rigorous way, it doesn't focus on what users need — an answer to "what is this?" The number of identified specimens will far exceed the number of type specimens, and many types will not be easily sequenced. Sequencing identified specimens puts the greatest amount of museum-based information into sequence space. This will become even more relevant as citizen science starts to expand to include DNA sequences (e.g., using tools like MinION).

Reality: Lack of clarity over what taxa to prioritise, emphasis on type specimens, concerns over whether DNA barcoding is out of date compared to other techniques (ignoring importance of global standardisation as a way to make data maximally useful) will all contribute to a piecemeal approach.

Who can help: Explore initiatives such as the Planetary Biodiversity Mission.

Vision: A physical visitor to the museum has a digital experience deeply informed by the museum's knowledge

Task: The physical walls of the museum are not barriers separating displays from science but rather interfaces to that knowledge. Any specimen on display is linked to what we know about it. If there is a fossil on a wall, we can instantly see the drawings made of that specimen in various publications, 3D scans to interact with, information about the species, the people who did the work (whether historical figures or current staff), and external media (e.g., BBC programs).

Reality: Piecemeal, short-lived gimmicky experiments (such as virtual reality), no clear attempt to link to knowledge that visitors can learn from or create themselves. Augmented reality is arguably more interesting, but without connections to knowledge it is a gimmick.

Who could help: Many of the links between specimens, species, and people full into the domain of Wikipedia and Wikidata, hence lots of opportunities for working with GLAM Wiki community.

Vision: A museum researcher can access all published information about a species, specimen, or locality via a single web site.

Task: All books and journals in the museum library that are not available online should be digitised. This should focus on materials post 1923 as pre-1923 is being done by BHL. The initial goal is to provide its researchers with the best possible access to knowledge, the secondary goal is to open that up to the rest of the world. All digitised content should be available to researchers within the museum using a model similar to the Haithi Trust which manages content scanned by Google Books. The museum aggressively pursues permission to open as much of the digitised content up as it can, starting with its own books and journals. But it scans first, sorts out permissions later. For many uses, full access isn't necessarily needed, at least for discovery. For example, by indexing text for scientific names, specimen codes, and localities, researchers could quickly discover if a text is relevant, even if ultimately direct physically access is

the only possibility for reading it.

Reality: Piecemeal digitisation hampered by the chilling effects of copyright, combined with limited resources means the bulk of our scientific knowledge is hard to access. A lack of ambition means incremental digitisation, with most taxonomic research remaining inaccessible, and new research constrained by needing access to legacy works in physical form.

Who could help: Consider models such as Hathi, work with BHL and publishers to open up more content, and text mining researchers to help maximise use even for content that can't be opened up straight away.

Vision: The museum as a "connection machine" to augment knowledge

Task: While a museum can't compete in terms of digital volume, it can compete for richness and depth of linking. Given a user with a specimen, an image, a name, a place, how can the museum use its extensive knowledge base to augment that user's experience? By placing the thing in a broader context (based on links derived from image -> identity tools, sequence -> identity tools, names to entities e.g., species, people and places, and links between those entites) the museum can enhance our experience of that thing.

Reality: The goal of having everything linked together into a knowledge graph is often talked about, but generally fails to happen, partly because things rapidly descend into discussions about technology (most of which sucks), and

squabbling over identifiers and vocabularies. There is also a lack of clear drivers, other than “wouldn’t it be cool?”. Hence expect regular calls to link things together (e.g., Let’s rise up to unite taxonomy and technology), demos and proof of concept tools, but little concrete progress.

Who can help: The Wikidata community, initiatives such as (some of these are no longer alive but useful to investigate) Big Data Europe, BBC Things. The BBC’s defunct Wildlife Finder is an example of what can be achieved with fairly simple technology.

Summary

The fundamental challenge the museum faces is that it is analogue in an increasingly digital world. It cannot be, nor should it be, completely digital. For one thing it can’t compete, for another its physical collection, physical space, and human expertise are all aspects that make a museum unique. But it needs to engage with visitors that are digitally literate, it needs to integrate with the burgeoning digital knowledge being generated by both citizens and scientists, and it needs to provide its own researchers with the best possible access to the museum’s knowledge. Above all, it needs to have a clear vision of what “being digital means”.

References

1. Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. Biodiversity Informatics, 7(2). https://doi.org/10.17161/bi.v7i2.3991

2. Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep Learning for Image-Based Cassava Disease Detection. Frontiers in Plant Science, 8. https://doi.org/10.3389/fpls.2017.01852

3. Xingchao Peng, Baochen Sun, Karim Ali, Kate Saenko (2014) Learning Deep Object Detectors from 3D Models. https://arxiv.org/abs/1412.7122
